5) Multi-Head Attention & Residual Connections

a) Why Multi-Head Attention?

Transformers use **multi-head attention** to allow the model to learn **different types of relationships** in parallel.
Each head attends to different parts of the sequence (e.g., positional, semantic, syntactic patterns), providing richer and more comprehensive contextual understanding than a single head could.

b) Purpose of Add & Norm (Residual + LayerNorm):

1. **Residual Connection (Add):**
   Helps preserve the original input information and allows gradients to flow more easily during backpropagation, **reducing vanishing gradient issues** and **stabilizing training**.

2. **Layer Normalization (Norm):**
   Normalizes activations to maintain stable feature distributions across layers, leading to **faster convergence** and **more stable training**.

c) Example of a Linguistic Relation Captured by Different Heads:

Different attention heads can focus on distinct relationships — for example,
one head might learn **coreference** (linking “he” to “John”),
while another might capture **syntactic structure** (connecting a verb to its subject).